DOCUMENT RESUME

ED 430 030 TM 029 774

AUTHOR Schafer, William D.; Swanson, Gwenyth; Bene, Nancy; Newberry, George

TITLE Effects of Teacher Knowledge of Rubrics on Student Achievement in Four Content Areas.

SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.

PUB DATE 1999-04-00

NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

CONTRACT R305F60143-96

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Academic Achievement; Biology; Civics; Constructed Response; Educational Assessment; English; High School Students; High Schools; *Knowledge Level; Mathematics; Scoring; *Secondary School Teachers; Student Attitudes; Student Evaluation; *Teacher Role; Test Construction

IDENTIFIERS *Scoring Rubrics

ABSTRACT

The hypothesis that enhanced knowledge of assessment rubrics by teachers and thus by students results in improved student achievement was studied in the context of the development of mandatory high school assessments for the Maryland State Department of Education. Rubrics were under development to score constructed-response items in the content areas of English, biology, mathematics, and government. Each of the state's 24 local education agencies was represented for 2 or 3 content areas. Forty-six pairs of teachers provided data. Half (the experimental group), received training in the state's Core Learning Goals, instructional strategies, and the use of rubrics as instructional tools. The remainder (control group), did not receive training. Data analyzed were effect sizes by form by item type within teacher pair. Results provide some empirical support for instructional uses of rubrics. Support was strongest for constructed-response items in biology, but was also seen for selected-response and constructed-response items in algebra. Neither positive nor negative effects were seen for English or government. Whether positive effects remain at the same levels across different types of rubrics remains for additional study. An appendix contains the generic rubrics. (Contains two tables and eight references.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 430 030

# Effects of Teacher Knowledge of Rubrics on Student Achievement in Four Content Areas

William D. Schafer
University of Maryland, College Park and
Maryland State Department of Education

Gwenyth Swanson, Nancy Bené, and George Newberry
Maryland State Department of Education

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
William Schafer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The authors wish to thank especially the MSDE content specialists whose efforts were essential to the success of this project. They are Elaine Crawford and Cindy Hannon (algebra), Gary Hedges (biology), Mary Jo Comer (English), Diane Johnson (government), and Susan Oskin (Skills for Success).

Presented at the American Educational Research Association Convention on April 22, 1999 in Montreal, Canada.

This work was supported, in part, by grant R305F60143-96 from the U. S. Department of Education's Office of Educational Research and Improvement.

TM029774

BEST COPY AVAILABLE

# Effects of Teacher Knowledge of Rubrics on Student Achievement in Four Content Areas

Careful attention to specification of achievement targets is believed to have value for both teachers and students (Stiggins, 1997). Stiggins and Conklin (1992) suggest that students, to predict what they will be asked to do in the future, use assessment patterns of teachers. Thus, teacher assessments are thought to serve as the operational definition of achievement for students. In this view, the more clearly achievement goals are known in terms of assessment activities, the more they will be accurately understood by each student and the greater the consistency of those understandings among a group of students (Wiggins, 1998). Assessments also may influence student motivation (Brookhart, 1997).

Similarly, the teacher who understands achievement goals in assessment (i.e., behavioral) terms is thought to be better equipped than those who do not to design effective instructional experiences (Airasian, 1994). That teacher is also believed to be better able to explain learning goals to students through descriptions of the activities they will be expected to engage in, improving accuracy of student understandings of achievement goals and criteria (Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996).

Rubrics are tools that can provide a way to translate achievement into assessment terms. As used here, a rubric is a rating scale that consists of ordered categories, together with descriptions and exemplars, which are used to sort student-produced responses into levels of achievement.

The purpose of this research was to study the hypothesis that enhanced knowledge of assessment rubrics by teachers, and thus by students, results in improved student achievement. It was hypothesized that teachers and students who have better understandings about what is expected will make effective use of that information. Teachers will design instructional experiences that capitalize on criteria expressed in their rubrics and will describe to students the relevant characteristics of demonstrations of achievement. Students in turn will be able to participate in the evaluation of their own work

## CONTEXT

This research was one of a series of exploratory studies intended to be of use to the Maryland State Department of Education (MSDE) in designing mandatory high school assessments for the state. It had already been determined that certain Core Learning Goals would be assessed and three item formats would be used: selected-response, brief (5-10 minutes) constructed response, and extended (15-30 minutes) constructed response.

Generic rubrics were under development to score constructed-response items in each of four content areas to be tested: biology, English, mathematics, and government. MSDE and local experts decided to develop generic rubrics, in part in order to facilitate their use in instruction by teachers and students. Several teacher committees had used the rubrics with student-produced responses to many items in attempts to select exemplars of the rubrics' score values, to learn

from the attempts, and to revise the rubrics.

The intent of MSDE, in this study, was to use all three item formats in each of three test forms. In addition, generic scoring rubrics would be used in instruction by one half of the study participants. These same rubrics would be used for scoring all constructed responses in the exploratory study. It was known that eventually, they would be modified for use with the high school assessment program. Thus, there was considerable motivation on the part of teachers and their supervisors to learn as much as possible about these aspects of the assessments along with the nature of the Core Learning Goals.

## METHOD

### Participants

Several groups of persons were involved at various phases in this study. Most participated in developing materials. These groups are described in the Materials section. Others were teachers who were administered the experimental manipulations; they are described here.

All school districts in the state were invited to nominate pairs of teachers in each of the four content areas to participate. The pairs of teachers were to have students with the same ability level, similar teaching and training background, and classes with similar demographics. The teacher pairs were to be teaching the same course (algebra I, biology, English, or government) . After the pairs were selected, the two teachers in each pair were randomly assigned either to receive or not to receive rubric training, the experimental manipulation.

Initially, each of the 24 Local Education Agencies (LEAs) in the state was represented by three of the four content areas except for one, which was represented by two. Because a few teachers dropped out late in the study, ultimately four LEAs were represented by only two content areas.

In making selections of which content areas would be represented in which LEAs, factors such as ability levels of classes and number of schools were taken into consideration. Some LEAs nominated only higher ability classes or average ability classes and some nominated pairs of classes that were not matched in ability level (or we learned by talking with teachers that their nominated classes could not necessarily be categorized by ability). Since we wanted our samples to be representative of all ability levels (about 50% of average ability and the rest evenly divided between higher and lower), in the content areas we were unable to include in the study those LEAs that did not provide an appropriate selection of student abilities. Additionally, we wanted matched classes to come from different schools that were comparable in demographics, rather than from the same school. In some of the smaller LEAs, that was not possible, and we used teachers with matched classes from the same school so that the LEA could be included in the study.

After selections and manipulations had been accomplished, but before tests were administered, some teachers withdrew from the study for reasons that ranged from scheduling conflicts to long term absences for surgical procedures to lack of continuing motivation. This produced a few

single, unmatched classes in each of the content areas. Although we administered the tests to these students, we did not include them in the analyses. There were 46 pairs of teachers who provided complete data. Approximately two-thirds of these 92 teachers were female and approximately 85% were white, not of Hispanic origin, the remainder African-American except for one teacher. About two-thirds of the teachers had between one and five years of teaching experience. Less than five percent were first-year teachers and less than five percent had more than 30 years of experience. The rest were about evenly distributed in five-year ranges from 6-10 years through 26-30 years.

In algebra, 12 (71%) of the initial 17 pairs of teachers remained at the end of the study. These teachers administered tests to 775 students. The average class size was 32.3.

Of the 17 initial pairs of biology teachers, 11 (65%) remained at the end. There were 838 students in their classes on the day of testing, and the average class size was 38.1.

There were 17 pairs of English teachers originally. At the end of the study, 11 pairs remained and administered tests to 760 students, for an average class size of 34.5.

In government, there were initially 20 pairs of teachers, of whom 12 (60%) remained at the end. These teachers administered tests to 818 students and the average class size was 34.1.

## Materials

### Core Learning Goals

The Core Learning Goals (CLGs) are intended to be the essential skills and knowledge that should be expected of Maryland high school graduates. They were developed by content teams and reviewed by educators and the public at large in order to guide local curricula throughout the state. The Maryland High School Core Learning Goals for English mathematics science and social studies define the body of knowledge upon which Maryland's eventual High School Assessment exams will be based. They have been approved by the State Board of Education.

The Core Learning Goal documents for each of the four content areas are organized into Goals, Expectations, and Indicators of learning. Goals represent the broad areas of content that students need to master. Expectations identify more defined topics or concepts within the goal areas. Indicators of learning address specific details of content that, when assessed, demonstrate student mastery.

A fifth Core Learning Goal area, called Skills for Success, contains learning, thinking, communication, interpersonal, and technology skills. Skills for Success are intended to define what high school students need to learn in addition to the knowledge and skills identified in their academic subjects. Skills for Success attempts to represent a focus that is general rather than subject-specific, applies equally well to all subjects, describes a context for the use of academic knowledge and skills, and provides tools for learning in any subject or skill area. They will be infused into each of the High School Assessments, but not assessed or scored separately.

## Rubrics in Each Content Area

A generic rubric was developed in each of the content areas for use in instructional activities and in scoring student-constructed responses. The same rubric was used for brief as well as extended responses. When teachers were trained in the use of rubrics, exemplars from actual student responses to both types of constructed responses were incorporated into the rubrics to illustrate the criteria as described in each of the score points.

Rubric development occurred at MSDE from the late summer of 1997 through early 1998 in each content area. Based primarily on Maryland's experience with statewide testing programs, especially the Maryland Writing Test, the content specialists and the high school scoring lead decided early in the process to develop generic rather than item-specific rubrics. Driving this decision was the idea of establishing goals that could be used in instruction and have a direct link to the scoring of the high school assessments.

Teams of four to six teachers and content specialists from the LEAs brainstormed criteria for development of rubrics that would be appropriate for the High School Assessments. The rubric writers used the draft rubrics in their classrooms to test the rubrics and harvest sample student responses to illustrate the qualities of the criterion that the score point descriptors contained. The rubric writers and other educators then reconvened to select, through consensus and independent scoring, appropriate exemplars and to refine the rubrics and verify that they were appropriate. From this work, 5-point scoring rubrics (0 to 4) were developed in each of the content areas. These appear in Appendix A

## Test Development

Three test forms were developed in each of the four content areas. Each form consisted of five selected-response (SR) items (multiple-choice with four options), two brief constructed response (BCR) items, and one extended constructed response (ECR) item.

In order to develop the tests, each MSDE content specialist identified six teachers to participate as item writers for the study. The initial item writing training occurred during February 1998, for two full days. Participants were brought together to receive training in thinking skills, Skills for Success, construction of open-ended questions, nature and use of scoring rubrics, Maryland's rubrics, Maryland's Core Learning Goals (presented in small groups by MSDE content area specialists), construction of selected response items, and characteristics of effective vs. ineffective items.

Prior to the workshop, each content specialist had identified Indicators from the Core Learning Goals that would reasonably be taught in or by April.

In algebra, the chosen content included recognition, description, and extension of patterns; representing patterns and functional relationships in tables, graphs, and mathematical expressions; adding, subtracting, multiplying, and dividing algebraic expressions; and description of graphs of non-linear functions including maxima, minima, roots, limits, rate of change, and continuity.

In biology, the content of the instructional unit included characteristics of chemical compounds and macromolecules utilized by living systems; structures of cellular and multicellular organisms; the identification and transmission of genetic traits; relationships between abiotic factors and biotic diversity; and changes in environmental conditions and their effects on the dynamics of populations.

In English, the content included composition of written presentations that inform, persuade, and express personal ideas; use of prewriting, drafting, and revision strategies of effective writers and speakers; and location, retrieval, and use of information from various sources to accomplish a purpose.

In government, the identified content included understanding the structure and functions of government and politics in the United States; evaluation of how the U. S. government has maintained balance between protecting rights and maintaining order; and explaining the influence of demographic changes on government policies.

Each team was responsible for entering March with a usable pool of items covering its content. Item writers had the flexibility to work independently or in small groups. Much of the work had to be done outside of the workshop on the participants' own time.

A second item development session occurred in March for one day at a system central office site. Participants included the item writers who had received training in February and a new team of item reviewers. The five members of each item review team included a teacher, an instructional supervisor, a representative for Skills for Success, a special educator, and a teacher of ESOL.

Item reviewers were asked to examine the pool of questions and to provide feedback in the form of recommendations about the accuracy of content, the appropriateness of the items, the accessibility of items to various student populations, etc. MSDE content specialists facilitated the group process. Using the recommendations provided by this critique, item writers revised questions to incorporate suggestions, ultimately to improve the quality of the item pool.

From this work, three similar test forms in each content area were constructed for administration in May. Each test form contained five selected response (SR) items and three constructed response (CR) items. Two of the CR items were brief (BCRs), expected to take about five minutes of student response time each, and one was extended (ECR), expected to take about 30 minutes of student response time. The forms were not developed to vary systematically and were considered interchangeable. In three of the content areas, the forms consisted entirely of non-overlapping items. In biology, there were only four BCRs that were judged acceptable and they were rotated such that each pair of forms had exactly one BCR in common.

Table 1 shows the alpha homogeneity reliabilities of each test form for each format and content. Homogeneity reliabilities for the selected-response item sets ranged from .14 to .67 and for the constructed-response item sets ranged from .31 to .75. There were five selected-response items on each form and they were designed to require about five minutes of testing time in all. The constructed-response items were expected to require about 40 minutes. For ease of comparison, the reliability estimates were re-estimated for a common one-hour (60-minute) time period in the Table using the Spearman-Brown formula. After testing, one BCR was deleted from one of the English forms (Form 2) because the scorers noted that very many students had clearly misunderstood the item.

## Procedures

The experimental manipulations occurred during March, 1998 at a local community college for two days. A first day of workshop training was attended by all teachers who administered the study (described in the Participants section) and the second day was attended by only those who received another day-long workshop on use of rubrics. A total of 71 pairs of teachers participated in the initial training.

One teacher from each pair was randomly assigned by the flip of a coin to group A, the other to group B. Group A teachers received training in Core Learning Goals, instructional strategies, and the use of rubrics as instructional tools. They became the experimental group. Group B teachers received the same training in Core Learning Goals and instructional strategies, but were excused for the day of rubric training sessions, and became the control group. Group A teachers were asked to incorporate the scoring rubrics created in Phase I into their daily lesson plans, utilizing them for instruction with their students. All teachers were told they would be administering tests provided by MSDE and were told the specific content from the Core Learning Goals that the tests would cover.

### Core Learning Goals Training

Teachers from both groups (A and B) were brought together at a local community college to receive training that emphasized the Indicators that would be tested within their respective content area Core Learning Goals and that suggested instructional strategies focusing on the implications of the Core Learning Goals for instruction. This was done for three reasons. First, we did not want instruction in general teaching techniques or familiarity with the Core Learning Goals to differ systematically between the two groups. Providing an introductory training session also allowed the later training session for the experimental group to focus on rubrics and their uses as opposed to instruction in general. Third, all participants received at least some attention through this training session. Although the general emphasis was the same in all content areas, each content area approached this training from a slightly different perspective.

In algebra, the training session focused on functions and algebra. Two content specialists worked with the participants to describe and then to list related implications for instruction.

After discussion, the Indicators were reviewed with the participants to verify that their understandings were accurate. For each content Indicator, sample instructional activities that addressed patterns and functional relationships in mathematics were provided. Additionally, test items that were appropriate to assessing those Indicators were suggested. Many of the instructional activities and assessment items came from textbooks widely used across the state. Instructional activities were selected to provide different representations of the same material using the language of mathematics and appropriate technology.

In the post-workshop evaluation, algebra participants felt that seeing problems selected specifically to evaluate student learning, the hands-on activities in the content area, and the opportunities to discuss ideas about teaching with colleagues were the most helpful aspects of the training session. The session was rated as 4 or higher (on a five-point scale) by 46% of the participants.

The biology training session was focused primarily on lesson design. It began by emphasizing that all lessons should be planned and that all test items should be developed making a deliberate attempt to combine science skills and process Indicators with actual biology content. The content specialist asked the participants to examine the language of the Indicators and to discuss the biology content limits. Participants were introduced to a "5-E" format of lesson planning (Engagement, Exploration, Explanation, Extension, Evaluation) and the value of connecting science Goals, Expectations, and Indicators to real world applications as well as to hands-on, inquiry-based instruction.

In the post-workshop evaluation, biology participants felt that discussions with peers about their attempts to incorporate all of the content Expectations and Indicators into the curriculum, the information provided about the use of the 5-E strategy for planning lessons, and the opportunity to network with other teachers were the most useful aspects of the training session. The session was rated as 4 or higher (on a five-point scale) by 71% of the participants.

In English, the training session focused just on the Indicators that would be tested. The content specialist discussed with participants the pertinent Goals, Expectations, and Indicators to provide an overview and clarification of the content they should cover. Participants then brainstormed instructional activities appropriate for those content Indicators and how, in general, to link instructional strategies to specific Indicators of learning. Further discussions were held in small groups.

In a post-workshop evaluation, English teachers commented that the large group session, the opportunity to share ideas with peers, and the ideas provided for connecting Indicators to instructional strategies were the most useful aspects of the training session. The session was rated as 4 or higher (on a five-point scale) by 57% of the participants.

Participants in the government Core Learning Goals session focused on political systems. The content and Skills for Success (SFS) specialists prepared model lessons for the social studies teachers combining SFS Indicators in thinking skills and communication skills with government Expectations. In small groups, participants were then asked to select a pair of Indicators (one

from SFS and one from government) and to prepare their own model lesson combining the two. Model lessons were shared with the large group, and critiqued feedback was encouraged.

In the post-workshop evaluation, government participants felt that learning to incorporate the Skills for Success with content, the opportunity to share strategies and ideas, and the clear explanation of the government Core Learning Goals were the most useful aspects of the training session. The session was rated as 4 or higher (on a five-point scale) by 69% of the participants.

Overall, 82% of participants agreed that their knowledge and understanding of the Core Learning Goals in their content area increased, and 77% agreed that their knowledge about instructional strategies and their uses increased.

### Rubric Training

Only the experimental group attended the second day of the training. Because we did not have available a workshop that we felt would not introduce irrelevant and confounding differences between the two groups, the control group was simply excused.

The session began with a large-group presentation. It included an overview of the characteristics of rubrics, how they can be used as a vehicle for conceptualizing student achievement for a teacher, and how they may be used in instruction to help show students how levels of their own achievement can be demonstrated and recognized and improved.

The rest of the day was spent in content groups led by content specialists. In each group, the generic rubric was distributed and discussed. They were then given anchor papers that had been developed along with the rubric and discrepancies between the score points and teacher judgments about the anchor papers were discussed. The teachers then scored training sets of student responses and discussed the application of the rubric and exemplars to assigning an accurate score. Finally, some teaching activities that incorporate rubrics were discussed. These included using examples to help students conceptualize levels of achievement according to the rubric, analyzing with students the language of the rubric and thinking with them about criteria they should use to assign score ratings to examples, peer review in which students revise each others' work, pairing students whose scores differ by only one point to discuss the differences, and the use of feedback forms that are tied to the rubric. Other uses were brainstormed in each group.

### Test Administration and Scoring

Content area exams were administered in May in algebra, biology, English, and government to both groups of students. Each teacher administered the three forms to random thirds of their classes. The papers were randomly ordered prior to forwarding to the teachers, and the teachers were instructed to distribute the exams to their students in the order they were received.

Selected response items were scanned at MSDE. Rangefinding, to select anchor responses for each constructed response consistent with the exemplars used in rubric training, occurred at

MSDE by teams of content area teachers and specialists, many of whom were involved in rubric construction. Qualified teams of readers scored student responses for the constructed response items. This was done by MSDE's commercial scoring contractor with participation of the MSDE content and scoring specialists during the summer.

Scorers used scoring guides produced during rangefinding to perform the ratings. Each scoring guide for an item consisted of a series of student responses chosen to illustrate the score points on the rubric. Each response is accompanied by a brief description of why it was scored at that rating. As part of training, each scorer participated in practice scoring by rating and discussing training sets prepared from actual students' responses.

Scoring was blind; no information identified papers as coming from the experimental group or the control group. Two readers, whose scores were averaged for the item score for that paper scored each CR response. If the readers differed by more than one score point, their scores were ignored and a new and highly experienced reader assigned the score for that item. Adjacent or exact agreement was reached in algebra for 99%, 99%, and 98% of the papers for the three test forms. In biology, the agreement rates for the forms were 97%, 99%, and 97%. The rates for English were 98%, 98%, and 97%. In government, the rates were 98%, 95%, and 91%. The rates are listed in the order of the forms in Table 1. To reflect the longer response time, the score for the ECR was doubled in computing a final CR score for each student. Thus, selected-response (SR) scores could vary from zero to five and constructed-response (CR) scores could vary from zero to 16.

## RESULTS

The primary analysis method was meta-analysis (Hedges & Olkin, 1985). The data that were analyzed consisted of effect sizes by form by item type within teacher pair. Each effect size was the mean of the students whose teacher received rubric training minus the mean of the students whose teacher did not, with the difference divided by the pooled within-groups standard deviation, and then adjusted for bias due to sample size. Any one pair of teachers produced an effect size for each of two item types (constructed response and selected response) for each of three forms (administered to a random third of each class of students), for a total of six effect sizes. In English and biology, we had available six effect sizes from each of 11 pairs of teachers, yielding 66 in each content, and totaling 132 from those two contents. From algebra and government there were six effect sizes for each of 12 teacher pairs, totaling 72 effect sizes for each content, and 144 in both. The grand total of effect sizes was 276.

The existence of several effect sizes from each teacher pair introduces dependencies into the analysis that violate the assumption of meta-analysis that the effect sizes are independent. One way to compensate for dependencies is to apply Bonferroni control over the familywise Type I error rate. In this analysis, the teacher pairs are independent, but there are six effect sizes from each pair. Bonferroni control was applied using the adjusted alpha of .05/6 = .008, and 99.2% confidence intervals are provided. It should be noted that no statistical significance decision would have been different had the .05 alpha level been used instead.

The meta-analysis yielded an overall estimate of effect size by pooling the 276 effect size estimates into one average. In this study, the overall effect size average was .1166. The standard error of the overall effect size was .0342 and the 99.2% confidence interval was .0265 to .2067. The confidence interval does not include zero and therefore the overall average effect size of .1166 is significantly different from zero at the .992 level of alpha.

The meta-analysis also yielded a chi-square statistic that evaluates whether there is evidence of heterogeneity of effect sizes about the overall average. The chi-square for heterogeneity was statistically significant ($\chi^2 = 515.309$, $df = 275$, $p < .001$).

Since the effect sizes varied significantly about their average, the effect sizes for the content areas were grouped and effect size averages were found separately. It was found that the four effect sizes differed significantly from each other ($\chi^2 = 20.766$, $df = 3$, $p < .001$). The pooled heterogeneity chi-square was also statistically significant, indicating that there was further variation that remains to be explained around the average content area effect sizes ($\chi^2 = 494.543$, $df = 272$, $p < .001$).

In algebra, the average effect size was .2868. Its standard error was .0672 and the 99.2% confidence interval was .1095 to .4641. The interval does not span zero and therefore the average effect size is significantly different from zero.

In biology, the average effect size was .2507 and its standard error was .0682. The 99.2 confidence interval was .0705 to .4309. Since the interval does not include zero, the average effect size is significantly different from zero.

In English, the average effect size was -.0318. Its standard error was .0731 and the 99.2% confidence interval was -.2248 to .1612. This interval does span zero and therefore the average effect size is not significantly different from zero. The 95% confidence interval was -.1752 to .1116, showing that the lack of statistical significance is not the result of low power (a wide confidence interval) due to use of the Bonferroni adjustment to alpha.

In government, the average effect size was -.0485 and its standard error was .0653. The 99.2% confidence interval was -.2209 to .1239. This interval includes zero and so the average effect size is not significantly different from zero. Since the 95% confidence interval is -.1766 to .0796, the lack of statistical significance is not the result of use of the Bonferroni alpha level.

Because the heterogeneity chi-square was significant, further modeling to differentiate the two formats, selected response (SR) and constructed response (CR), was undertaken within each content. These results are discussed below. Following separation of the effect sizes into the eight groups by content and format, the combined heterogeneity chi-square remained statistically significant ($\chi^2 = 487.862$, $df = 271$, $p < .001$). No further modeling was undertaken since we had exhausted the variables available to us to predict effect size. The results for each of the combinations of content area by item type appear in Table 2.

Within algebra, the SR effect size was .2656, the standard error was .0948, and the confidence

interval was .0155 to .5157. The CR effect size was .3082, the standard error was .0953, and the confidence interval was .0567 to .5597. Neither of these intervals span zero and therefore the effect size for both item types is significantly different from zero. The difference between the effect sizes for the formats was not statistically significant ($\chi^2 = .110$, $df = 1$, $p = .739$).

Within biology, the SR effect size was .0498, the standard error was .0954 and the confidence interval was -.2020 to .3016. The CR effect size was .4615, the standard error was .0978, and the confidence interval was .2035 to .7195. The SR confidence interval spans zero but the CR interval does not. Therefore, only the CR effect size is significantly different from zero. The 95% interval for the SR effect size is -.1373 to .2369, so the lack of statistical significance is not due to the use of the Bonferroni adjustment. The difference between the effect sizes for the formats was statistically significant ($\chi^2 = 9.081$, $df = 1$, $p < .003$).

Within English, the SR effect size was -.0686, the standard error was .1031, and the confidence interval was -.3406 to .2034. For the CR forms, the effect size was .0054, the standard error was .1038, and the confidence interval was -.2684 to .2793. Both span zero and so neither is significantly different from zero. The 95% interval for the SR effect size was -.2707 to .1335 and for the CR effect size was -.1980 to .2089. Thus, neither would have been significantly different from zero had the Bonferroni adjustment not been applied. The difference between the effect sizes for the formats was not statistically significant ($\chi^2 = .257$, $df = 1$, $p = .612$).

The SR effect size within government was -.1343, the standard error was .0928, and the confidence interval was -.3792 to .1106. For government, the CR effect size was .0358, the standard error was .0920, and the confidence interval was -.2069 to .2785. Had the Bonferroni adjustment not been applied, the 95% intervals would have been -.3162 to .0476 for the SR effect size and -.1445 to .2161 for the CR effect size, neither being significantly different from zero. The difference between the formats was not statistically significant ($\chi^2 = 1.705$, $df = 1$, $p = .192$).

While the individual format-content results are the best to interpret, it should be mentioned for completeness that the overall difference between the format effect sizes across contents was not statistically significant ($\chi^2 = 6.492$, $df = 1$, $p < .011$), although it would have been significant if the Bonferroni adjustment not been applied. The SR effect size across contents was .0301 and the confidence interval was -.0969 to .1572. This interval spans zero and therefore the SR effect size is not significantly different from zero. The CR effect size was .2042 and its confidence interval was .0763 to .3321, which is significantly different from zero because it does not span zero. The 95% SR effect size confidence interval would have been -.0643 to .1245 had the Bonferroni adjustment not been applied, so the lack of significance was not due to low power for the purpose of familywise Type I error rate control.

An alternative analysis could have been to conduct repeated-measures analysis of variance separately by contents using the class means, either for forms separately or summed across forms. That approach was not taken because it would have been impossible to compare the item formats or the contents directly, an advantage of the meta-analysis.

## DISCUSSION

The results of this study provide some empirical support for instructional uses of rubrics. That support was strongest for constructed response items in biology, but also appeared for both selected response and constructed response formats in algebra. There was no evidence found to support instructional uses of rubrics in English or in government, but neither was there support for a position that the use of rubrics has detrimental effects on achievement.

Since there were differences noted between the content areas, as well as the item formats, an evaluation of the results should treat them separately. In algebra, the overall effect size was approximately 0.3. Although the effect size for the constructed response items was greater than that for the selected response items, they were both in the same general range. Indeed, it can be argued from the algebra results that knowledge of the rubric to be used in evaluation of student work produces a learning advantage that generalizes across assessment formats.

No support for rubric effects was found for any item type in either English or government. One possible explanation is that teachers in these areas already use generic rubrics, either explicitly or implicitly, in their grading, and perhaps even instructionally. If so, the training experience may not have contained much, if any new information for those teachers. That may be especially true in the area of English. A writing test has been a graduation requirement in Maryland for many years and it is scored using rubrics that, in the judgment of English specialists at MSDE, are not very different from the generic rubric used here with the English teachers.

A multiple-choice test of citizenship skills had been a requirement for graduation for students studying government up to and including the year that this study was conducted. That test is based on a highly specific curriculum that does not emphasize the higher-order thinking embodied in the Core Learning Goals in government. It is possible that at least some government teachers felt emphasizing an expanded set of content goals and higher-order thinking processes would be a disservice to their students who needed to be successful on the citizenship skills test. At lease some of these teachers may therefore have ignored, with respect to their classroom activities, both the Core Learning Goals and the rubric understandings provided in this study.

The results for biology suggest a positive effect size of 0.5 for constructed response items but no effect was supported for selected response items. The magnitude of the difference for constructed response items was the largest found. It is possible that the use of a generic rubric is very different from the way biology teachers typically grade student constructed responses, which may be more focused on presence or absence of specific elements idiosyncratic to the task.

Since the clearest results in favor of the rubric group were found in algebra, it is worthwhile to examine the rubric for algebra in comparison with those for the other content areas. There are five elements that appear throughout the four non-zero score values. These are application, representation, explanation, justification, and analysis. Each of these may be capable of being described to a student more easily than the terminology used in the other content area rubrics, such as understanding, extending meaning, tone, and insight. It is possible that the latter terms,

whose meanings may require greater elaboration, are less useful than those in the mathematics rubric for instructional purposes are. The degree of abstraction in the terminology used to express rubrics could be varied systematically in future research.

In order to interpret the magnitudes of the effect sizes found, it should be recalled that the effect size in this study represents the advantage, in standard deviation units, of a one-day educational experience about rubrics for a group of teachers. Interpretation of the magnitude of the effect size index was discussed by Cohen (1988), who gave guidelines of 0.2 for "small" effects, 0.5 for "medium" effects, and 0.8 for "large" effects.

A small effect size of 0.2 is the sort of magnitude Cohen (1988) felt would occur in a new field of inquiry. Lack of understandings that are needed in developing strong treatments was offered as one of the reasons effect sizes would be so small in a beginning behavioral science area. The current study focused only on understandings about rubrics. It did not manipulate knowledge of the objectives. Indeed, all teachers received the same presentation on the goals of instruction (the Core Learning Goals). Further, the nature of these rubrics as generic, while common in some fields, such as writing, is unusual in other disciplines. The developers of the treatments in this study to enhance knowledge and uses of the rubrics by teachers had little research to guide them. Given this lack of direction from research on the effects of rubrics, effect sizes of 0.2 should be expected.

It is possible that improved rubrics or stronger treatments could be designed so that rubric training might be found effective in all content areas. Indeed, the rubrics used in this study were drafts that have since been revised. It is also possible that the positive effects of rubric training are already common within the repertoire of teachers in certain disciplines. But the results of this study are encouraging of the use of rubrics as one way to promote more effective instruction by teachers. In order to do that, understandings about the ways teachers may use rubrics in classrooms and the ways they are perceived by students are needed. It may then be possible to design effective ways to enable teachers to become skilled in their instructional uses of rubrics. For example, it may be that the advantages of understandings about rubrics become enhanced over longer periods of time, which would provide opportunities for teachers to learn how to use rubrics and to internalize those uses.

While this study has demonstrated, in at least some content areas, positive effects of focused understandings about rubrics, whether those effects remain at the same levels across different types of rubric, such as analytic, or objective-specific, remains for additional study. It will also be useful to study whether positive effects can be found to generalize across assessment formats, such as in the algebra results reported here, since that would suggest the advantages of rubrics in the learning process are demonstrable irrespective of the way in which learning is demonstrated. Further research should also be undertaken to explore the mechanisms by which instructional change due to rubrics is effected as well as the nature of the resulting student behaviors that lead to improvements in achievement.

References

Airasian, P. W. (1994). Classroom assessment (2nd ed.). New York: McGraw-Hill.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. Applied Measurement in Education, 10(2), 161-180.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Hedges, L. V. & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. Educational Measurement: Issues and Practice, 15(3), 7-18.

Stiggins, R. J. (1997). Student-centered classroom assessment (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Stiggins, R. J. & Conklin, N. F. (1992). In teachers' hands: Investigating the practices of classroom assessment. Albany: SUNY Press.

Wiggins, G. (1998). Educative assessment: Designing assessments to inform and improve student performance. San Francisco: Jossey-Bass.

Table 1.

Internal Homogeneity of the Tests across Students

| Content | Format | Form | Alpha | One-Hour Alpha |
|---|---|---|---|---|
| Algebra | | | | |
| | Selected | | | |
| | | 1 | .52 | .93 |
| | | 2 | .35 | .87 |
| | | 3 | .37 | .88 |
| | Constructed | | | |
| | | 1 | .56 | .66 |
| | | 2 | .73 | .80 |
| | | 3 | .62 | .71 |
| Biology | | | | |
| | Selected | | | |
| | | 1 | .16 | .70 |
| | | 2 | .23 | .79 |
| | | 3 | .29 | .83 |
| | Constructed | | | |
| | | 1 | .63 | .72 |
| | | 2 | .68 | .76 |
| | | 3 | .53 | .63 |
| English | | | | |
| | Selected | | | |
| | | 1 | .67 | .96 |
| | | 2 | .48 | .92 |
| | | 3 | .34 | .86 |
| | Constructed | | | |
| | | 1 | .69 | .77 |
| | | 2* | .31 | .44 |
| | | 3 | .69 | .77 |
| Government | | | | |
| | Selected | | | |
| | | 1 | .32 | .85 |
| | | 2 | .14 | .66 |
| | | 3 | .42 | .90 |
| | Constructed | | | |
| | | 1 | .70 | .78 |
| | | 2 | .75 | .82 |
| | | 3 | .64 | .72 |

*One brief constructed response item was deleted from this test during scoring.

Table 2.

Effect Sizes and Confidence Intervals by Content Area and Item Format

| Content | Format[b] | Average Effect Size | Interval Limits[a] Lower | Upper |
|---|---|---|---|---|
| Algebra | SR | .2656 | .0155 | .5157 |
| | CR | .3082 | .0567 | .5597 |
| Biology | SR | .0498 | -.2020 | .3016 |
| | CR | .4615 | .2035 | .7195 |
| English | SR | -.0686 | -.3406 | .2034 |
| | CR | .0054 | -.2684 | .2793 |
| Government | SR | -.1343 | -.3792 | .0476 |
| | CR | .0358 | -.2069 | .2785 |

[a]The confidence coefficient is 99.2% for all intervals.
[b]SR represents selected response (multiple-choice)items and CR represents constructed response (essay) items.

Appendix A: Generic Rubrics

Algebra Rubric

4. A response at this level analyzes the full range of the problem correctly. It represents all of the information appropriately, and applies mathematical concepts, which are essentially complete and correct, to solve the problem. This response thoroughly explains the process(es) used to solve the problem and justifies clearly the conclusion(s) in the context of the problem. This response must have the correct answer.

3. A response at this level analyzes the problem correctly. It represents most of the information appropriately, and applies mathematical concepts to solve the problem correctly. It may contain minor flaws. This response explains the process(es) used and shows some justification of the conclusion in the context of the problem.

2. A response at this level analyzes most of the problem correctly. It represents some of the information appropriately, and applies mathematical concepts to solve the problem. It may contain major flaws, or be incomplete. This response shows little or no attempt to explain the process used or to justify the conclusion.

1. A response at this level shows some attempt to solve the problem but analyzes the problem incorrectly. It represents little or no information appropriately and makes some attempt to apply mathematical concepts to solve the problem. This response makes little or no attempt to explain the process used or to justify the conclusion. It may have the correct answer with no supporting information or have inappropriate mathematical concepts.

0. A response at this level shows no evidence of mathematical thinking or no response is given.

Justify conclusion means the student will use mathematical principles (definitions, postulates, theorems) to support the reasoning used to solve the problem.

Explain the processes used means the student will use the language of mathematics to communicate how the student arrived at the answer.

A major flaw is an error that affects solving the problem.

A minor flaw is an error that does not affect solving the problem.

Biology Rubric

4. There is evidence in this response that the student, using analysis, has developed a full and complete understanding of the question or problem. The student has synthesized information to provide a correct answer and the supporting evidence demonstrates an integration of ideas. Information provided in the response demonstrates that the student has extended scientific concepts beyond the question or problem. The response is enhanced through the use of accurate terminology to explain scientific principles.

3. There is evidence in this response that the student, using analysis, has developed a good understanding of the question or problem. The student has synthesized information to provide a correct answer and the supporting evidence is complete. The response uses mostly accurate terminology to explain scientific principles.

2. There is evidence in this response that the student has a basic understanding of the question or problem. The student provides a correct answer, but the supporting evidence is only moderately effective. The response uses limited accurate terminology to explain scientific principles.

1. There is evidence in this response that the student has some understanding of the question or problem. The student may provide a correct answer, but the supporting evidence is only minimally effective. The response makes little use of accurate terminology to explain scientific principles.

0. There is evidence that the student has no understanding of the question or problem. The student provides an incorrect answer or does not answer the question. The response does not make use of scientific terminology.

English Rubric

4. The response demonstrates a thoughtful creation, examination, and extension of meaning expressed in a distinctive voice and a deliberate tone. Through development, it fulfills a purpose and provides relevant support and complete elaboration. It organization skillfully follows an established structure that enhances the purpose throughout. Control of language demonstrates conscious selection of words and careful attention to audience understanding and interest and to conventions.

3. The response demonstrates the creation, examination, and extension of meaning and maintains a consistent voice and tone. Through development, it establishes a purpose and uses support and elaboration. Its organization adequately follows a structure and supports the purpose with minor inconsistencies. Choice of language demonstrates attention to audience understanding and interest and to conventions with few, if any, errors which interfere with meaning.

2. This response shows some creation, examination, and extension of meaning. Although development may be incomplete, the response addresses a purpose and uses support and elaboration. Its organization may lack consistency and therefore may not fully support the purpose. Use of language suggests some attention to audience understanding and interest and to conventions. The response may contain errors which interfere with meaning.

1. The response shows an attempt to create, examine, and extend meaning. Its development may include either a purpose that is not evident or may be misunderstood or contains support or elaboration that is inadequate. Its organization may be unstructured and confusing. Use of language may show minimal attention to audience understanding and interest and to conventions. The response may contain errors which interfere with meaning.

0. Other.

Government Rubric

4. This response shows understanding of the historical development and/or current status of principles, institutions, or processes of political systems. The response is insightful, integrates knowledge, and demonstrates powerful application.

3. This response shows some understanding of the historical development and/or current status of principles, institutions, or processes of political systems. The response is thorough with accurate documentation and appropriate application.

2. This response shows knowledge of the historical development and/or current status of principles, institutions, or processes of political systems. The response is acceptable with some support or key ideas. The response makes some use of higher order thinking skills.

1. This response shows minimal knowledge of the historical development and/or current status of principles, institutions, or processes of political systems. The response is related to the question but is inadequate, with significant misconceptions and/or absence of key ideas.

0. This response is unrelated to the question or does not provide an answer to the question.